SegSLR: Promptable Video Segmentation for Isolated Sign Language Recognition

Sven Schreiber[0009−0009−8652−9549], Noha Sarhan[0000−0002−1545−9346], Simone Frintrop[0000−0002−9475−3593], and Christian Wilms[0009−0003−2490−7029]

Computer Vision Group, University of Hamburg, Germany
firstname.lastname@uni-hamburg.de

Abstract. Isolated Sign Language Recognition (ISLR) approaches primarily rely on RGB data or signer pose information. However, combining these modalities often results in the loss of crucial details, such as hand shape and orientation, due to imprecise representations like bounding boxes. Therefore, we propose the ISLR system SegSLR, which combines RGB and pose information through promptable zero-shot video segmentation. Given the rough localization of the hands and the signer's body from pose information, we segment the respective parts through the video to maintain all relevant shape information. Subsequently, the segmentations focus the processing of the RGB data on the most relevant body parts for ISLR. This effectively combines RGB and pose information. Our evaluation on the complex ChaLearn249 IsoGD dataset shows that SegSLR outperforms state-of-the-art methods. Furthermore, ablation studies indicate that SegSLR strongly benefits from focusing on the signer's body and hands, justifying our design choices.

Keywords: Sign language recognition · Action recognition · Segmentation.

1 Introduction

Sign language is a central way to communicate for the deaf or hard-of-hearing. It transmits information through several visual parameters, most importantly manual parameters like hand shape, orientation, position, or movement, but also non-manual parameters like body posture, facial expressions, or head movements [1]. Outside the deaf or hard-of-hearing community, few people understand sign language, which substantially limits the social interaction of the deaf or hard-of-hearing. To bridge this gap, Isolated Sign Language Recognition (ISLR) systems classify a video sequence of sign language on a gloss-level.

ISLR systems rely on various techniques [25,30]. Several systems explicitly focus on the most relevant image areas like the signer's body, hands, or face to capture the manual and non-manual parameters [7,12,4,32,31,13,18]. This is done by attending to the relevant parts of the RGB video frames (RGB-based) or adding dedicated networks, which encode pose information (pose-based). Pose-based models extract keypoints of the signer and subsequently generate a skeleton-like

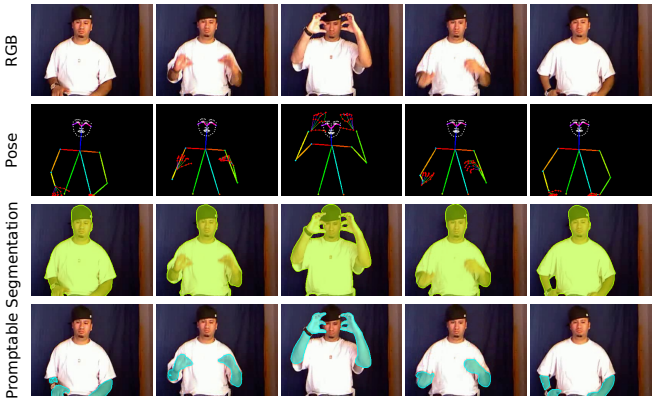


Fig. 1: Idea of the proposed SegSLR system: We combine RGB information (first row) and pose information (second row) through promptable video segmentation. The pose information is used as prompts to segment the RGB frames. This leads to segmentations of the signer's body (third row) and hands (fourth row) to focus processing on the most relevant image areas for ISLR.

graph, which is then processed by complex architectures [22,13,18,33]. In contrast, RGB-based models mostly extract crops around the signer's hands or the signer itself. These crops are classified as one part of the system. Note that some RGB-based methods use pose information to locate the relevant parts in the image. Thus, most methods either operate solely on RGB or pose information, or use the pose mostly to guide the extraction of crops, losing critical details about the pose, such as hand shape, hand orientation, or body posture.

Recently, foundation models for promptable segmentation of images and videos revolutionized segmentation in several domains and tasks [3,37,24,9,35]. Promptable segmentation refers to segmenting a coherent part of the image given a suitable prompt, like point coordinates or a bounding box. SAM [16] for images and SAM 2 [26] for videos, address these tasks and generate high-quality segmentations in a zero-shot manner without domain-specific training data. Hence, these models are also well-suited for ISLR.

In this paper, we combine RGB-based and pose-based ISLR through the innovative use of promptable video segmentation in our novel ISLR system SegSLR, visualized in Fig. 2. SegSLR uses a multi-stream approach. Besides Inflated 3D CNN (I3D CNN) [2] based streams for plain RGB information and optical flow, ensuring a comprehensive understanding of both spatial and temporal dynamics, SegSLR contains segmentation streams, which combine RGB and pose information through promptable video segmentation. The innovative idea of these streams is, first, to utilize pose information to locate the signer's body and hands in the form of point sets. Second, these point sets are utilized by the promptable video segmentation method SAM 2 to generate high-quality segmentations (masklets) of the signer's body and hands through the video in

a zero-shot manner to retain rich pose information (see Fig. 1 for an example). Finally, these masklets are used to focus the processing of the RGB frames in the subsequent classifiers based on I3D CNNs to the most important image areas for ISLR. This results in a unified architecture exclusively relying on simple I3D CNNs for classification. Based on this innovative combination of pose and RGB information, SegSLR outperforms state-of-the-art systems on the complex ChaLearn249 IsoGD dataset, as our evaluation shows. Ablation studies also demonstrate the benefit of the key design decisions.

Overall, our contributions are three-fold:

- We introduce SegSLR, a novel ISLR method, leveraging both RGB and pose information.
- We use foundation models for promptable video segmentation in ISLR to combine pose and RGB information, which retains rich pose information through high-quality segmentations.
- SegSLR outperforms the existing state-of-the-art by up to 4.17%, while ablation studies validate the design choices.

2 Related Work

Methods for ISLR evolved from models based on hand-crafted features to learned CNN-, LSTM-, or transformer-based architectures [25,30]. Here, we will discuss learned methods since they substantially outperform traditional ones. Learned models for ISLR can be divided into three groups: methods that mainly rely on RGB information, on pose information, or hybrid methods.

Methods relying only on RGB input either encode the entire frame or, additionally, crops of relevant areas. [28] encode entire RGB frames of a video in an I3D CNN architecture [2] and add a second stream with optical flow to better capture the temporal dynamics. Subsequently, the streams are fused on score-level. Similarly, [20] apply I3D CNN to the RGB input and add weakly supervised data to improve the feature extraction. [41] combine 3D CNNs for short-term temporal dependencies with a Conv LSTM for long-term ones.

Several RGB-based methods additionally focus on the signer's body parts as important parameters for sign language recognition. [21] apply simple hand detectors based on Haar-like features to extract bounding boxes around the hands and track them through the video, yielding a hand energy image for classification. [32,29,31] all extend the two-stream model of [28] by focusing on different aspects. While [32] extract masks of the hands in an individual stream to mask the RGB input, [29] focus the RGB input on the moving areas in an end-to-end learned manner, resulting in attending hands and arms. Finally, [31] utilize pseudo depth as an additional stream, effectively segmenting the signer without combining this information with the RGB stream.

Using only pose information, [22] apply a graph convolutional network (GCN) to the skeleton graph generated by a pose estimation system, while also applying advanced data augmentations, which mix signs. Also utilizing GCNs as their backbone, [36] add text embeddings in a contrastive learning framework, while

[33] use GCNs at the frame-level and model the temporal dependency using BERT [5]. [13,40] use BERT for self-supervised pre-training. In [13], masked hand poses are reconstructed for pre-training. In contrast to previous methods, [18] apply transformer-based modules instead of GCNs.

Hybrid methods combine RGB and pose information. [14] propose a multi-stream architecture with 3D CNNs for RGB and optical flow input as well as a GCN for the skeleton graph, combining them through score fusion. Several systems [7,12,8,4] crop parts of the RGB frames or latent representations around the hands, face, or the entire signer for focused processing based on pose information. Yet, only bounding boxes [7,12,4] or rough segmentations [8] are used. Finally, [42] encode the position of specific joints of the signer into heatmaps that can be processed by 3D CNNs, which also capture RGB information [14].

Closest to our proposed SegSLR are the hybrid methods, which use pose information to focus processing of the RGB data on relevant areas like the signer's body or hands. However, [7,12,4] only use bounding boxes, losing information about the hand shape, hand orientation, or the signer's body posture, which are all highly relevant for sign language recognition. [8] generate pixel-precise segmentations, yet they are derived from the pose information through dilation. This results in low-quality segmentation masks. In contrast, SegSLR employs high-quality segmentations of the signer's body and hands using a foundation model, retaining important details about body posture and manual parameters.

3 Revisit Segment Anything Model 2

Original SAM [16] for images is a segmentation system trained on a large-scale dataset, which is applied to various tasks in a zero-shot manner. To guide SAM, prompts (points, boxes, or masks) are used, which indicate what to segment within an image. SAM 2 [26] extends this idea to videos and leverages the temporal information to segment consistent masklets. SAM 2 consists of three major components: encoders, a decoder, and a memory mechanism. The image and prompt encoders consist of a masked autoencoder [10] for images as well as positional and learned embeddings for the prompts. Given these embeddings, the mask decoder predicts a segmentation mask and an IoU score. To ensure temporal consistency across the video, a memory attention mechanism is added. For training SAM 2, [26] proposed a new dataset SA-V, with manually or semi-automatically annotated videos.

4 Method

This section introduces our novel SegSLR system for ISLR, as depicted in Fig. 2. SegSLR follows the general approach of [28], comprising a multi-stream architecture based on Inflated 3D CNNs (I3D CNNs) [2] with streams for processing plain RGB frames and derived optical flow information, as visible in the center and top of Fig. 2. In addition, SegSLR has four segmentation streams (see bottom part in Fig. 2) to effectively combine RGB and pose information. To this

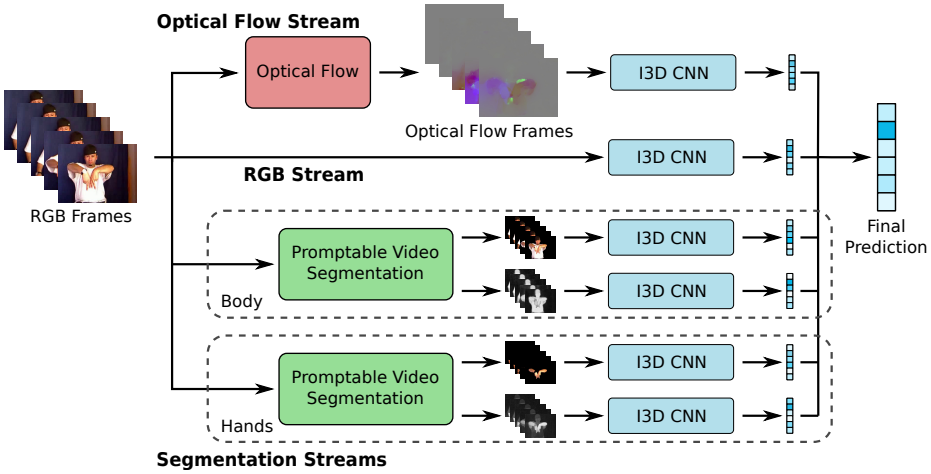
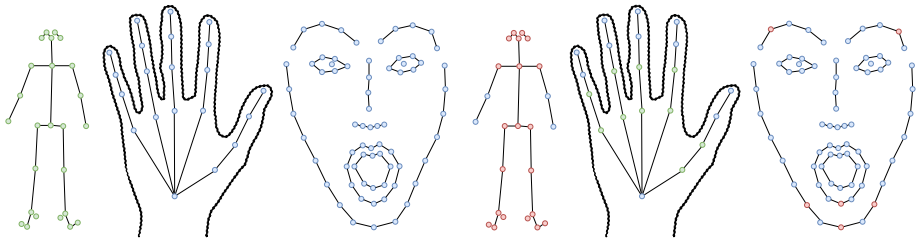


Fig. 2: Overview of our proposed SegSLR system. Based on RGB frames of a video, the first stream (Optical Flow Stream) calculates the optical flow and, subsequently, classifies these optical flow frames using an I3D CNN. The second stream, RGB Stream, directly applies an I3D CNN to the plain RGB frames. As the key novelty of SegSLR, we propose the segmentation streams, which combine RGB and pose information using promptable video segmentation modules (see Fig. 4) to generate four outputs: RGB frames focused on the signer's body and hands as well as the respective segmentation logits. These outputs are processed by I3D CNNs. Finally, score-level fusion is applied to aggregate the results.

end, a promptable video segmentation module as visualized in Fig. 4 first estimates the signer's pose and generates keypoints for the signer's body and hands. A selection of these keypoints is used to prompt SAM 2, yielding high-quality segmentation masks of the signer's body and hands. Subsequently, two streams in SegSLR directly take the logits of the segmentation masks for the body and both hands, respectively. The other two segmentation streams mask the RGB frames, to focus processing on the relevant parts of the RGB input while maintaining rich pose information (hand shape, hand orientation, and body posture) through the high-quality segmentations. Finally, the data of each stream is processed by an I3D CNN per stream, and the results of all I3D CNNs are combined using score-level fusion (see right part in Fig. 2). In the following sections, we discuss each step in more detail.

4.1 Pose Estimation and Prompt Generation

As a first step in all segmentation streams of SegSLR, we derive point prompts based on pose information for prompting SAM 2 in our promptable video segmentation module (see Fig. 4). As outlined in Sec. 1, pose estimation results are used in ISLR to capture the body posture or locate the hands [7,12,8,4]. Therefore, we take the state-of-the-art human pose estimation system RTMW [15]



(a) Selected keypoints for the body (b) Selected keypoints for the hands

Fig. 3: Overview of RTMW [15] keypoints for prompting SAM 2 to segment the signer's body (a) and hands (b). Green keypoints are positive point prompts, red keypoints are negative point prompts, and blue keypoints are ignored.

and initially extract 116 keypoints, as visualized in Fig. 3, covering the entire signer's body. Given these keypoints, we create two subsets. The first subset (see Fig. 3a) captures the entire signer's body, therefore, we use all keypoints except for detailed hand and face keypoints, since this level of detail is not necessary. The second subset focuses on the hands and includes the keypoints around the first joint per finger, as visualized in Fig. 3b. We ignore the other hand keypoints since they are frequently outside the hands due to imprecise localization. We also select negative keypoints to discern the hands from the remaining body in SAM-2. As negative keypoints, we select all major body keypoints and important face keypoints. All keypoints in both subsets are the point prompts to guide SAM 2. Note that undetected keypoints are ignored for prompt generation.

4.2 Best Frame Selection

Before applying SAM 2, we select the best frame to start the segmentation through the video in our promptable video segmentation module. Frame selection is important since the first frame might not cover the hands, as the signer's hands are usually at hip level. For instance, this is apparent from the initial frame in Fig. 1. To select the best frame, we assess the quality of the detected keypoints per frame by calculating the average keypoint confidence, the size of the bounding box around all keypoints, and the overlap between hands and face keypoints. The keypoint confidence is a per-keypoint output from RTMW and a surrogate for RTMW's confidence about the keypoint. To calculate a single score, we take the average across all detected keypoints. The remaining two measures are used to select a frame, which shows no or only a minimal overlap between the hands and the face. Hands in front of the face frequently occur in sign language. However, due to their similarity in color, it is challenging to discern these regions for a segmentation system when they overlap. Therefore, we first calculate the area of the bounding box covering all keypoints. This helps to measure how sprawled the arms are. To prevent cases where only the upper arms are stretched out and the hands are still close to the face, we determine

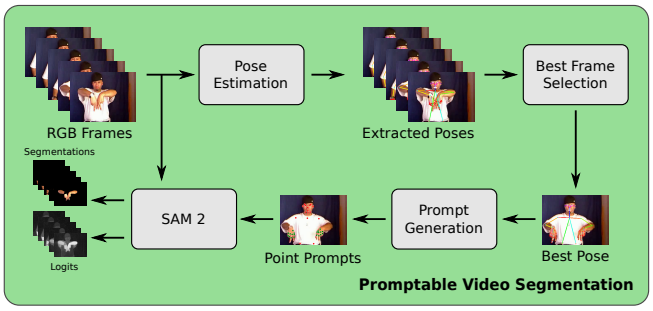


Fig. 4: Detailed view of our promptable video segmentation module for hands segmentation. For segmenting the signer's body, the same pipeline is applied with different point prompts. From the input RGB frames, the signer's pose is estimated, and the best frame to initiate the video segmentation is determined. In this frame, several keypoints are converted to positive or negative point prompts. Using these prompts, SAM 2 is applied to the RGB frames. The results are the RGB frames masked by the SAM 2 segmentations and the segmentations' logits. Note that the estimated poses and the point prompts are overlayed with the RGB frames for visualisation only.

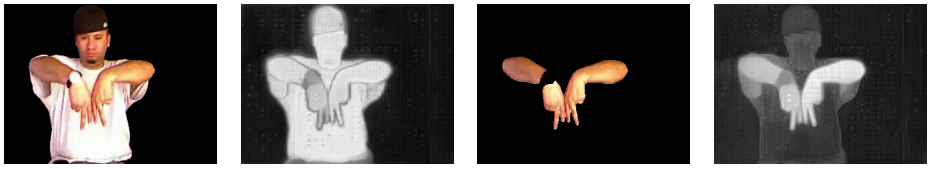
the maximum overlap between one hand's bounding box and the bounding box around the face. Note that all bounding boxes are created based on the detected RTMW keypoints. To combine the three scores, we first normalize the scores by their respective per-video maximum, subtract the maximum overlap from 1, and multiply them. Hence, the best frame will receive the maximum score. Given this frame, we use the respective sets of point prompts for prompting SAM 2.

4.3 Mask Generation

Given the two sets of point prompts for the signer's body and hands on the best frame of a video sequence, we bidirectionally prompt SAM 2 starting from the best frame in our promptable video segmentation module. Hence, we apply SAM 2 from the best frame forward through the video and backward. This process is conducted for each set. The resulting masklet per set is applied to the RGB frames to focus on either the signer's body or hands. Besides the binary segmentation masks, we also extract the per-frame logits of the segmentations to capture the global scene context through SAM 2's per-pixel confidence.

4.4 ISLR Classification

Our new segmentation streams are integrated into the framework of [28] for ISLR classification. This results in SegSLR, as visible in Fig. 2. SegSLR consists of three main streams, one for the plain RGB data, one for optical flow data extracted from the RGB data using [38], and the aforementioned segmentation



(a) Body segmentation (b) Body logits (c) Hands segmentation (d) Hands logits

Fig. 5: Examples of the body segmentation, body logits, hands segmentation, and hands logits as intermediate outputs of the segmentation streams in SegSLR.

streams. As visible from Fig. 2, there exist four segmentation streams working on the outputs of the two promptable video segmentation modules for segmenting the signer's body and hands. The first stream uses masked RGB frames, effectively masking the background and focusing the processing on the signer based on the pose information. In contrast, the second stream processes the logits of the body segmentation. Similarly, for the two hands, the third and fourth streams do the same, processing masked RGB frames and the logits for both hands simultaneously. This focuses processing on the dominant parameter for ISLR, the hands. In case of the logits, the captured global context includes information about the location of the signer, making the relative location of the hands visible (see Fig. 5d). In contrast to most ISLR works focusing on hands, the precise and temporally consistent segmentations by SAM 2 capture the hand shapes, hand orientation, and even details about the fingers as visible in the examples in Fig. 5. We evaluate the design choices regarding the streams in Sec. 5.2.

Finally, each segmentation stream in SegSLR uses an I3D CNN as commonly applied in ISLR literature [20,28,12] to capture spatial and temporal dependencies. The per-stream results are combined using score-level fusion.

4.5 Implementation Details

SegSLR consists of three trainable components: (1) SAM 2, (2) RTMW, and (3) six I3D CNNs. For SAM 2, we apply a model pre-trained on SA-1B [16] and SA-V [26] as proposed by [26] and do not add fine-tuning, since segmentation annotations are unavailable in ISLR datasets. Similarly, we directly apply the pose estimation system RTMW, which is pre-trained on 14 datasets as suggested by [15]. For the I3D CNNs in SegSLR, we follow [28] and pre-train the I3D CNNs on ImageNet [27] and the Kinetics dataset [2]. Subsequently, each I3D CNN is fine-tuned individually with Adam optimizer, a batch size of 4, and early stopping with a patience of 3. We apply standard categorical cross-entropy loss per stream. When training or testing on videos, we uniformly sample 40 frames and extract a central 224×224 crop from each frame. This ensures a constant size and length of the videos for the I3D CNNs. We additionally utilize data augmentation for training the I3D CNNs and shift the extracted crop horizontally or vertically and adjust the brightness [28].

5 Evaluation

We evaluate SegSLR on the commonly used ChaLearn249 IsoGD dataset [34] comprising 249 gestures across 47,933 videos in complex environments and under challenging lighting conditions. For details about the covered gestures, we refer to [34]. We use the training dataset to train SegSLR and report results on the validation and test sets. Note that we train SegSLR five times on the training set and report the result with the median validation accuracy for a fair comparison. This model is also used to generate the test results. We additionally report the mean and standard deviation across the five trainings. We compare SegSLR against several recent ISLR methods utilizing RGB and/or pose information. This includes the baseline system of SegSLR [28], methods that focus on the signer's body [31,19] or hands [32,19], and a method combining RGB and pose information [19]. Note that we do not compare to methods utilizing depth, since this would be unfair, given the strong semantic cues of depth data for ISLR. A comparison to other methods like [7,12,8] is impossible due to missing publicly available implementations. For assessing the quality, we use standard accuracy.

5.1 Results on ChaLearn249 IsoGD Dataset

Table 1 presents the results on the validation and test sets of ChaLearn249 IsoGD. The results clearly show that SegSLR outperforms all other methods on both sets by a substantial margin. This includes outperforming methods, which focus on the signer's hands through bounding boxes [32,19] or the signer's body [31,19]. The earlier explicitly shows the advantage of utilizing hand segmentations preserving hand shape details over simple boxes. Moreover, [19] also combine RGB and pose information by focusing their system on boxes around the hands, elbows, and shoulders of the signer. Comparing SegSLR to its baseline system, I3D-SLR, which only comprises the RGB stream and the optical flow stream, SegSLR shows an improvement of 9.21% and 8.32% in accuracy on validation and test sets. This is the result of adding the segmentation stream, combining RGB and pose information. Compared to the other variations of I3D-SLR, adding pseudo depth [31], focusing on moving areas [29], and focusing on the signer's hands [32], SegSLR shows large improvements of up to 8.80% and 6.56% on the validation and test sets, respectively. Overall, this shows the strong performance of SegSLR based on combining pose and RGB information through promptable video segmentation.

The qualitative segmentation results in Fig. 6 on ChaLearn249 IsoGD support the aforementioned quantitative results of SegSLR. Across all videos in Fig. 6, the segmentations of the signer's body and hands are accurate and suitable to focus the processing of SegSLR on these highly relevant areas. Even despite substantial hand movement or highly-textured backgrounds, the segmentations consistently capture the signer's body and hands. Specifically, in the videos in the first two rows, a waving hand is visible that challenges SAM 2 due to the high velocity of the movement. Yet, the segmentations are consistent. In the second and third rows, the videos show glosses where a hand is moved

Table 1: Comparison of the proposed SegSLR to state-of-the-art ISLR methods on the ChaLearn249 IsoGD dataset. For SegSLR, we report the median accuracy on the validation set over five trainings, along with the respective mean and standard deviation in parentheses. *: Numbers were not reported by the respective paper.

Method	**Accuracy** (%)	
	Validation	**Test**
C3D-LSTM [41]	43.88	-*
SYSU ISEE [19]	50.02	-*
XDETVP [39]	51.31	-*
8-MFFs-3flc (5 crop) [17]	57.40	-*
I3D-SLR [28]	62.09	64.44
I3D-pseudoDepth [31]	62.50	66.20
2SCVN-RGB-Fusion [6]	62.72	-*
Hybrid Attn-I3D-SLR [29]	65.02	68.89
TD-SLR [32]	67.13	70.91
SegSLR (ours)	**71.30** ($\mu = 71.39$, $\sigma = 0.41$)	**72.76**

Table 2: Results of SegSLR on the ChaLearn249 IsoGD dataset with three versions for the segmentation streams. Body_{RGB} denotes the segmentation stream utilizing masked RGB frames. Body_{Logits} is the segmentation stream processing the raw logits of the body segmentation. Hands_{RGB} and Hands_{Logits} are the respective streams segmenting the hands.

Input Streams	**Accuracy** (%)	
	Validation	**Test**
Base	62.09	64.42
+ Body_{RGB}	65.51	67.33
+ Body_{RGB} + Body_{Logits}	67.10	69.78
+ Body_{RGB} + Body_{Logits} + Hands_{RGB} + Hands_{Logits}	**71.30**	**72.76**

in front of the face. Despite visual similarity between the hands and the face, SAM 2, prompted with our point prompts derived from pose information, distinguishes between hands and face. This is due to the negative prompts covering the face when generating the hand segmentation. Finally, the last row also shows an example of a sign with a complex finger posture. Still, the hand segmentation in SegSLR captures all fingers in detail, given the pose-based point prompts.

5.2 Ablation Studies

Input Streams To assess the impact of SegSLR's focus on the signer's body and hands, and the importance of utilizing both the binary segmentation mask and the logits, we present in Tab. 2 the step-by-step results, from the baseline system [28] to SegSLR. In each step, we add segmentation streams to combine

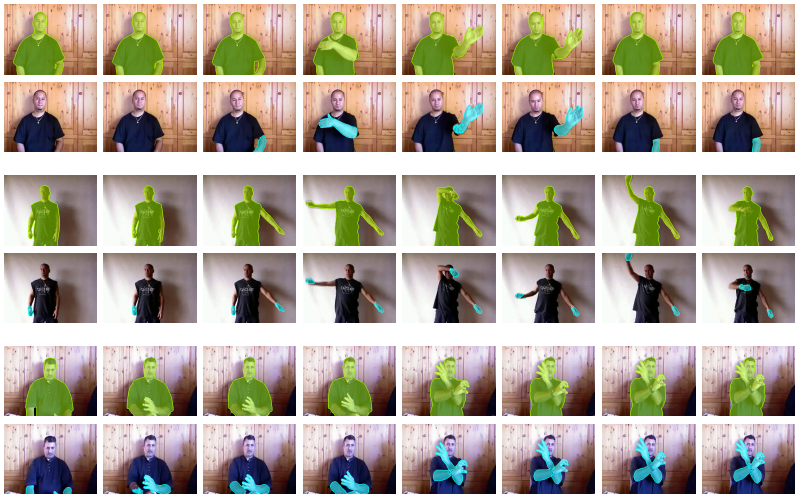


Fig. 6: Qualitative body (green masks) and hands (blue masks) segmentation results generated within our SegSLR on videos of the ChaLearn249 IsoGD test dataset. For each video, one sequence of frames is overlayed with the body segmentation and the hands segmentations, respectively. Note that we only show 8 of the 40 frames per video for brevity.

pose and RGB information. Given the baseline, we first add SegSLR's body segmentation by masking the respective RGB frame (Body_{RGB} in Tab. 2). The improvements of 3.42% and 2.91% clearly indicate the advantage of an additional focus on the signer's body. Adding the logits of the body segmentation in another stream (Body_{Logits} in Tab. 2) further improves the results by 1.59%/2.45% and shows the additional value of the logits. Finally, adding the same segmentation streams for the hands (Hands_{RGB} and Hands_{Logits} in Tab. 2), further improves the results by 4.22% and 2.98%, highlighting the importance of hands for ISLR. Overall, the results in Tab. 2 clearly show the value of all segmentation streams combining RGB and pose information.

Segmentation Method We present the results of SegSLR with three different segmentation methods: well-known Mask R-CNN [11] trained on the COCO dataset [23] to segment only humans, denoted as Mask $\text{R-CNN}_{\text{Person}}$, SAM [16], and SAM 2 [26] as in the proposed SegSLR. Note that we utilize only the segmentation stream for the signer's body in SegSLR to match Mask R-CNN's training on COCO. For SAM and SAM 2, we apply the same keypoints as described in Sec. 4.1 The results in Tab. 3 show that SegSLR outperforms the baseline with any segmentation stream. Yet, SAM and SAM 2 surpass Mask R-CNN by up to 3.20%, which underlines the strong zero-shot segmentation ability of these foundation models. Comparing SAM and SAM 2 shows that SAM 2 leads to an improvement of 1.37% and 0.87%. A major reason for this is the temporal

Table 3: Results of SegSLR with different segmentation methods on the ChaLearn249 IsoGD dataset. Note that SegSLR only includes the body segmentation stream (Body_{RGB}) here.

Input Streams	**Accuracy** (%) **Validation**	**Test**
Base	62.09	64.42
Base + Mask R-$\text{CNN}_{\text{Person}}$	62.31	65.44
Base + SAM (Body_{RGB} only)	64.14	66.46
Base + SAM 2 (Body_{RGB} only)	65.51	67.33

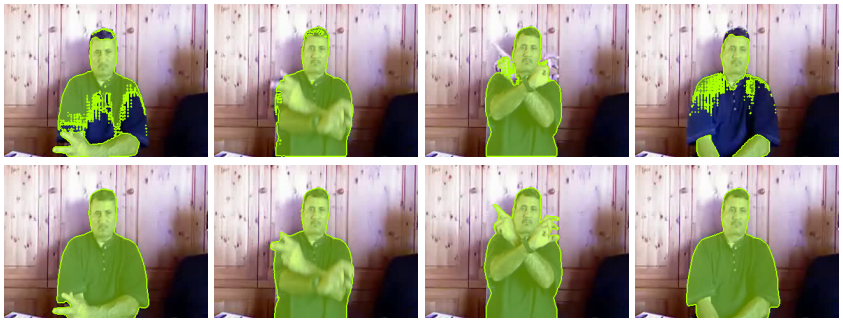


Fig. 7: Comparison of segmentations of the signer's body with SAM (upper row) and SAM 2 (lower row) in SegSLR on four frames of a video from the ChaLearn249 IsoGD test dataset.

consistency of the segmentations between the frames. This is also visible in the qualitative results in Fig. 7, where the SAM segmentations (upper row) flicker substantially between the frames, while the SAM 2 segmentations (lower row) consistently cover the entire body.

6 Conclusion

Effectively utilizing both RGB and pose information is key for high-quality ISLR. To address this combination of RGB and pose information without losing details important to understand sign language, we proposed the novel ISLR system SegSLR. It innovatively combines RGB and pose information through promptable video segmentation using pose keypoints to prompt SAM 2 and detect the signer's body and hands for a focused processing of the RGB data. The strength of this novel design for ISLR is supported by our experiments on the ChaLearn249 IsoGD dataset, where SegSLR outperforms all competing methods. Our ablation studies also validated the focus of SegSLR on the signer's body and hands as well as the use of SAM 2 based on prompts from pose estimation data. Overall, SegSLR presents another step to bridge the communication gap between people inside and outside the deaf or hard-of-hearing community.

References

1. Baker-Shenk, C.L., Cokely, D.: American Sign Language: A teacher's resource text on grammar and culture. Gallaudet University Press (1991)
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Conference on Computer Vision and Pattern Recognition (2017)
3. Cheng, H.K., Oh, S.W., Price, B., Schwing, A., Lee, J.Y.: Tracking anything with decoupled video segmentation. In: International Conference on Computer Vision (2023)
4. De Coster, M., Van Herreweghe, M., Dambre, J.: Isolated sign recognition from RGB video using pose flow and self-attention. In: Computer Vision and Pattern Recognition Workshop (2021)
5. Devlin, J.: BERT pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Duan, J., Wan, J., Zhou, S., Guo, X., Li, S.Z.: A unified framework for multi-modal isolated gesture recognition. Transactions on Multimedia Computing, Communications, and Applications (2018)
7. Gökçe, Ç., Özdemir, O., Kındıroğlu, A.A., Akarun, L.: Score-level multi cue fusion for sign language recognition. In: European Conference on Computer Vision Workshop (2020)
8. Gruber, I., Krnoul, Z., Hrúz, M., Kanis, J., Bohacek, M.: Mutual support of data modalities in the task of sign language recognition. In: Computer Vision and Pattern Recognition Workshop (2021)
9. He, C., Li, K., Zhang, Y., Xu, G., Tang, L., Zhang, Y., Guo, Z., Li, X.: Weakly-supervised concealed object segmentation with SAM-based pseudo labeling and multi-scale feature grouping. In: Advances in Neural Information Processing Systems (2024)
10. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Conference on Computer Vision and Pattern Recognition (2022)
11. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask R-CNN. In: International Conference on Computer Vision (2017)
12. Hosain, A.A., Santhalingam, P.S., Pathak, P., Rangwala, H., Kosecka, J.: Hand pose guided 3D pooling for word-level sign language recognition. In: Winter Conference on Applications of Computer Vision (2021)
13. Hu, H., Zhao, W., Zhou, W., Wang, Y., Li, H.: SignBERT: Pre-training of hand-model-aware representation for sign language recognition. In: International Conference on Computer Vision (2021)
14. Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., Fu, Y.: Skeleton aware multi-modal sign language recognition. In: Computer Vision and Pattern Recognition Workshop (2021)
15. Jiang, T., Xie, X., Li, Y.: RTMW: Real-time multi-person 2d and 3d whole-body pose estimation. arXiv preprint arXiv:2407.08634 (2024)
16. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: International Conference on Computer Vision (2023)
17. Kopuklu, O., Kose, N., Rigoll, G.: Motion fused frames: Data level fusion strategy for hand gesture recognition. In: Computer Vision and Pattern Recognition Workshop (2018)

18. Lee, T., Oh, Y., Lee, K.M.: Human part-wise 3d motion context learning for sign language recognition. In: International Conference on Computer Vision (2023)
19. Li, B., Li, W., Tang, Y., Hu, J.F., Zheng, W.S.: GL-PAM RGB-D gesture recognition. In: International Conference on Image Processing (2018)
20. Li, D., Yu, X., Xu, C., Petersson, L., Li, H.: Transferring cross-domain knowledge for video sign language recognition. In: Computer Vision and Pattern Recognition (2020)
21. Lim, K.M., Tan, A.W.C., Lee, C.P., Tan, S.C.: Isolated sign language recognition using convolutional neural network hand modelling and hand energy image. Multimedia Tools and Applications **78** (2019)
22. Lin, K., Wang, X., Zhu, L., Zhang, B., Yang, Y.: SKIM: Skeleton-based isolated sign language recognition with part mixing. Transactions on Multimedia **26** (2024)
23. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision (2014)
24. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1) (2024)
25. Rastgoo, R., Kiani, K., Escalera, S.: Sign language recognition: A deep survey. Expert Systems with Applications (2021)
26. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
27. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115** (2015)
28. Sarhan, N., Frintrop, S.: Transfer learning of videos: From action recognition to sign language recognition. In: International Conference on Image Processing (2020)
29. Sarhan, N., Frintrop, S.: Sign, attend and tell: Spatial attention for sign language recognition. In: International Conference on Automatic Face and Gesture Recognition (2021)
30. Sarhan, N., Frintrop, S.: Unraveling a decade: A comprehensive survey on isolated sign language recognition. In: International Conference on Computer Vision Workshop (2023)
31. Sarhan, N., Willruth, J.M., Fritnrop, S.: Pseudodepth-slr: Generating depth data for sign language recognition. In: International Conference on Computer Vision Systems (2023)
32. Sarhan, N., Wilms, C., Closius, V., Brefeld, U., Frintrop, S.: Hands in focus: Sign language recognition via top-down attention. In: International Conference on Image Processing (2023)
33. Tunga, A., Nuthalapati, S.V., Wachs, J.: Pose-based sign language recognition using GCN and BERT. In: Winter Conference on Applications of Computer Vision (2021)
34. Wan, J., Zhao, Y., Zhou, S., Guyon, I., Escalera, S., Li, S.Z.: Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In: Computer Vision and Pattern Recognition Workshop (2016)
35. Wilms, C., Rolff, T., Hillemann, M., Johanson, R., Frintrop, S.: SOS: Segment object system for open-world instance segmentation with object priors. In: European Conference on Computer Vision (2024)
36. Wong, R., Camgoz, N.C., Bowden, R.: Learnt contrastive concept embeddings for sign recognition. In: Computer Vision and Pattern Recognition Workshop. pp. 1945–1954 (2023)

37. Wu, J., Ji, W., Liu, Y., Fu, H., Xu, M., Xu, Y., Jin, Y.: Medical SAM adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620 (2023)
38. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime tv-l1 optical flow. In: German Conference on Pattern Recognition (2007)
39. Zhang, L., Zhu, G., Shen, P., Song, J., Afaq Shah, S., Bennamoun, M.: Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition. In: International Conference on Computer Vision Workshop (2017)
40. Zhao, W., Hu, H., Zhou, W., Shi, J., Li, H.: BEST: BERT pre-training for sign language recognition with coupling tokenization. In: AAAI Conference on Artificial Intelligence (2023)
41. Zhu, G., Zhang, L., Shen, P., Song, J.: Multimodal gesture recognition using 3-d convolution and convolutional LSTM. IEEE Access **5** (2017)
42. Zuo, R., Wei, F., Mak, B.: Natural language-assisted sign language recognition. In: Computer Vision and Pattern Recognition (2023)